Yesterday I debated Eliezer Yudkowsky

This is totally me spitballing here, but I think the crux of the debate is whether AIs will look more like formal systems or like inscrutable weight matrices.

I think it's the latter, and I think it's the something like the latter forever. This is a good thing.

I'm so glad we reached an empirical point of disagreement. Can superintelligent AIs "solve" the one shot prisoners dilemma and play cooperate-cooperate? Yudkowsky thinks yes, I think no.

	C	D
C	(3,3)	(0,5)
D	(5,0)	(1,1)

The payoff matrix makes the solution very clear. You want to convince your opponent to cooperate while you defect. Of course you end up with defect-defect while you'd both rather cooperate-cooperate, but that's just how it works. Is this what the field of AI alignment is trying to "solve"? For two complex systems in the real world with time bounds, I'm 99% sure this is impossible.

MIRI has done research into this, and has shown it was possible for "modal agents." However, it's clearly not solvable for programs in general. I can't see how any practical AI agents would be able to prove properties like this about each other.

Also, that paper stipulates "algorithms with read-access to one anothers' source codes." If I'm negotiating with someone, the last thing I want to do is give them access to my source code!

Another point of disagreement was over AI's using formal programming languages like Coq or Lean. The reason humans don't use these languages much isn't due to a limitation of humans, it's due to program search being harder in those spaces.

Eventually computer systems will be better at programming than me. But this doesn't mean they won't ever write bugs! If you are writing 0 bugs, you are coding too slowly, and in a competitive enviroment you will lose.

Search and optimization are not magic, they are just hard. As we accept ASI won't be able to crack AES keys, we should also accept that many other kinds of search and optimization are still very hard for them, like one-shotting diamond nanobots!

The brain also seems to be pretty close to the Landauer limit. This makes sense, if there's one thing that evolution brutally optimized for it is efficiency.

Our current silicon is 10-10000x off the limit, here's a good talk about it. We'll be hitting the limit soon, and it's unclear if reversible computing or quantum computing will pan out.

Our brain is likely the size it is because that's optimal efficiency wise. Evolution could have made bigger brains, but they will use more power, and that just wasn't a good trade off for our ancestral environment. And if you are optimizing for inclusive genetic fitness, it's still unclear if it's a good trade off today!

I think the rise of the thinking machines will be slow and predictable, just like the rise of the muscle machines was. All the machines in the world have 2 zettaflops of compute. All the humans have 160,000 zettaflops. It's not clear which is working together better, consider that a wash. The machines still have 16 doublings to go to catch us.

We produce about 0.5 zettaflops of compute per year, so let's say the compute in the world doubles every 4 years (of course assuming the compute production keeps up in percent). The human record for population doubling was in the 60's at about every 35 years. The machines are reproducing about 10x faster.